GOLDX at SemEval-2025 Task 11: RoBERTa for Text-Based Emotion Detection

Bill Clinton
Institut Teknologi Bandung
summerrs528@gmail.com

Abstract

This document describes an approach to solve the SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, specifically for Track A: Multi-label Emotion Detection and Track C: Cross-lingual Emotion Detection. In this document, the method utilizes an ensemble of RoBERTa models, each trained with different hyperparameters to enhance robustness and performance. For Track C, an additional neural machine translation (NMT) approach is added. The results demonstrate the effectiveness of model ensembling and translation preprocessing in tackling the challenges posed by Task 11.

1 Introduction

The SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection focuses on multi-label emotion detection, which is a key challenge in the natural language processing (NLP) world. Its applications include sentiment analysis, user engagement analysis, and mental health monitoring. This task involves predicting the presence or absence of specific emotions (in the form of 0 or 1) for a given text. Track A requires detecting five different emotions: anger, fear, joy, sadness, and surprise, using a labeled dataset in a single language (for English). Track C, on the other hand, introduces a cross-lingual setting where no labeled training data is provided in the target language. Moreover, Track C includes an additional emotion category, which is disgust, increasing the total number of labels to six in certain languages. Participants must develop strategies to generalize across languages without access to direct supervision in the target language.

In this paper, the approach leverages an ensemble of RoBERTa models, each trained with different hyperparameters, to improve the robustness and performance. For Track A, the training is done using the provided labeled dataset (English). For Track C, a cross-lingual transfer learning strategy is adopted (training on Chinese data and inferring on Indonesian). Additionally, MarianMT, a neural machine translation model developed by Helsinki-NLP, is employed to translate non-English data before classification, helping bridge the linguistic gap.

Through this participation, strong results are achieved, obtaining a Macro-F1 score of 0.762 for Track A and 0.4291 for Track C. This system performed competitively relative to other submissions, demonstrating the effectiveness of model ensembling and translation-based adaptation.

2 System Description

2.1 Overview

In this paper, the approach is based on pretrained Transformer model from Hugging Face, specifically using the RoBERTa model for multi-label emotion classification. This system is designed to optimize macro-F1 scores for each emotion through hyperparameter tuning. An ensemble of RoBERTa models in implemented, each trained with different hyperparameters to improve performance.

For Track A, the training directly used the provided labeled dataset in English. For Track C, where no labeled target-language data is available, a cross-lingual approach is used:

- The model is trained on Chinese emotion-labeled data.
- The inference is conducted on Indonesian test data using the trained model

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 797–800
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Additionally, MarianMT (Helsinki-NLP) is leveraged for the translation preprocessing, ensuring consistency between training and test languages.

2.2 Model and Hyperparameter Tuning

To optimize the model's performance, experiments are done with multiple hyperparameter settings, selecting the best configuration based on macro-F1 scores per emotion. The following hyperparameters were tuned:

- Batch Size: {8, 16, 24, 28, 32, 40, 48, 54}
- Learning Rate: {2e-5, 5e-5, 1e-4}
- Epochs: {4, 5, 8, 10, 12, 16}

For each experiment on the test dataset in the development phase, the model achieving the highest macro-F1 score for each individual emotion was selected for inference.

2.3 Cross-Lingual Transfer in Track C

One of the main challenges in Track C is the lack of labeled training data in the target language. To address this, here is the approach:

- Chinese is used as the training language, leveraging its availability of labeled data
- MarianMT is applied to translate both the training and test data into English before being processed.
- The output from the ensemble models are aggregated to determine the final classification.

3 Data

3.1 Dataset Overview

The experiments in this paper utilized the BRIGHTER dataset collection, which is a comprehensive resource for multi-label emotion recognition across 28 languages, predominantly focuing on low-resource languages from regions, such as Africa, Asia, Latin America, and Eastern Europe. Each dataset comprises the text instances annotated by fluent speaker. This captured a diverse range of emotional expressions. The primary emotions annotated are anger, fear, joy, sadness, surprise, and disgust. Notably, the presence of the disgust label varies across languages. For instance, the label is included in the Chinese dataset but not in the English dataset.

3.2 Data Collection and Annotation

The data collection process involved sourcing text from various domains to ensure a rich and diverse representation of emotional expressions. Fluent speakers of each language were recruited to annotate the datasets. This ensured cultural and contextual relevant in the emotion labels. Annotators were provided with guidelines to label each text instance with one or more emotions, reflecting the multi-label nature of the task. This approach acknowledges the complexity and nuance of human emotions, where a single sentence can convey many emotions.

4 Experimental Setup

4.1 Data Splits and Usage

For both Track A and Track C, the dataset was split into train (80%) and dev (20%) from the labeled data. The train set was used for training the models, while the dev set was used for hyperparameter tuning.

4.2 Preprocessing

There is an extra preprocessing step for track C, which is using MarianTokenizer. This is used fro the translation step for both the Chinese training data and the Indonesian test data.

4.3 Model and Hyperparameter Tuning

This system utilized a model ensemble approach, where multiple versions of the same base model were trained with different hyperparameters. The best model for each emotion label was chosen based on its macro-F1 score on the dev set.

- Data Split for Model Training:
  - Labeled data split: 80% training, 20% development
  - Test set: Used only for final evaluation
- Hyperparameter Search:
  - Learning rate: {2e-5, 5e-5, 1e-4}
  - Batch size: {8, 16, 24, 28, 32, 40, 48, 54}

- Epochs: {4, 5, 8, 10, 12, 16}

4.4 Hardware and External Libraries

- Hardware: Experiments were conducted using an NVIDIA A100 GPU on Google Collab Pro
- External Libraries:
  - Hugging Face Transformers (transformers)
  - PyTorch (torch)
  - NumPy (numpy)
  - Pandas (pandas)

5 Results

The system is evaluated in two phases:

- Development Phase (Until January 16, 2025)
- Test Phase (Until February 1, 2025)

5.1 Track A Results

Emotion	Dev	Test
Anger	0.8	0.6817
Fear	0.8271	0.8488
Joy	0.8214	0.7732
Sadness	0.8056	0.7738
Surprise	0.7797	0.7324
Macro-F1	**0.8067**	**0.762**

Table 1: Track A Results.

5.2 Track C Results

Emotion	Dev	Test
Anger	0.5897	0.4884
Fear	0.1951	0.2589
Joy	0.5536	0.6392
Sadness	0.3243	0.5507
Surprise	0.3736	0.4797
Disgust	0.125	0.1575
Macro-F1	**0.3602**	**0.4291**

Table 2: Track C Results.

Selected Models:

- Model I:
  - Batch Size: 32
  - Learning Rate: 5e-5
  - Epoch: 5
  - Emotions: Anger
- Model II:
  - Batch Size: 40
  - Learning Rate: 5e-5
  - Epoch: 5
  - Emotions: Disgust, Fear, Joy, Surprise
- Model III:
  - Batch Size: 28
  - Learning Rate: 5e-5
  - Epoch: 4
  - Emotions: Sadness

The results indicate that the ensemble approach and cross-lingual transfer strategy are quite effective. Track C, despite its challenges, achieves a good performance, highlighting the benefits of multilingual training and translation-based inference.

References

FacebookAI. (2019). RoBERTa-large: A Robustly Optimized BERT Pretraining Approach. Hugging Face. Retrieved from https://huggingface.co/FacebookAI/roberta-large

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692.* Retrieved from https://arxiv.org/pdf/1907.11692.

Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Wahle, J. P., Ruas, T., Beloucif, M., de Kock, C., Surange, N., Teodorescu, D., Ahmad, I. S., Adelani, D. I., Aji, A. F., Ali, F. D. M. A., Alimova, I., Araujo, V., Babakov, N., Baes, N., Bucur, A.-M., Bukula, A., Cao, G., Cardenas, R. T., Chevi, R., Chukwuneke, C. I., Ciobotaru, A., Dementieva, D., Gadanya, M. S., Geislinger, R., Gipp, B., Hourrane, O., Ignat, O., Lawan, F. I., Mabuya, R., Mahendra, R., Marivate, V., Piper, A., Panchenko, A., Porto Ferreira, C. H., Protasov, V., Rutunda, S., Shrivastava, M., Udrea, A. C., Wanzare, L. D. A., Wu, S., Wunderlich, F. V., Zhafran, H. M., Zhang, T., Zhou, Y., & Mohammad, S. M. (2025). BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages**.** *arXiv preprint arXiv:2502.11926.* Retrieved from https://arxiv.org/abs/2502.11926.

Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Yimam, S. M., Wahle, J. P., Ruas, T., Beloucif, M., De Kock, C., Belay, T. D., Ahmad, I. S., Surange, N., Teodorescu, D., Adelani, D. I., Aji, A. F., Ali, F., Araujo, V., Ayele, A. A., Ignat, O., Panchenko, A., Zhou, Y., & Mohammad, S. M. (2025). SemEval Task 11: Bridging the Gap in Text-Based Emotion Detection. Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria. Association for Computational Linguistics.

Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.-A., Nieminen, T., Raganato, A., Scherrer, Y., Vázquez, R., & Virpioja, S. (2023). Democratizing neural machine translation with OPUS-MT. Language Resources and Evaluation, 58, 713–755. Springer Nature. https://doi.org/10.1007/s10579-023-09704-w

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT), Lisbon, Portugal.